AIGC 时代新闻舆论工作新阵地——面向大模型的可信训练数据集与服务能力建设

蔡津津
（新华社媒体融合生产与技术系统国家重点实验室，北京 100031）

【摘要】
【目的】生成式人工智能 AIGC 的出现和广泛应用对新闻舆论格局产生了颠覆性的影响，使算法和算力逐步进化成为高质量内容生产和传播的权力核心，新闻舆论工作需要在新趋势下掌握主动权。【方法】AIGC 大模型成为潜在的社会舆论成员并以远超人类个体的知识面和内容处理生成速度在潜移默化中掌握舆论引导的话语权，而决定 AIGC 大模型能力和价值观立场的核心是训练数据集的构建。【结果】随着美西方价值观和意识形态数据集训练下产生的 AIGC 大模型在全球的普及，我国主流新闻舆论工作面临着严峻挑战与风险，必须开辟面向大模型的可信训练数据集和数据服务的建设阵地。【结论】不仅可以做到“守土有责”，履行好议题设置、舆论引导、内容生产和传播的把关人角色，更可以通过规范准确、代表主流价值观和意识形态的数据集与服务供给，抢占 AIGC 时代舆论引导、思想引领、文化传承、服务人民的传播高地。
【关键词】AIGC，新闻舆论，ChatGPT，意识形态，训练数据集

一、导语

随着万物互联的下一代信息技术飞速发展，数字世界与现实世界的融合不断加深，大规模数据与算力共同推动的人工智能技术跨越式发展，全球科研团队都在致力于让人工智能具备人类理解、思考、逻辑推理和输出内容的能力，从而大幅降低人类操作数字世界来改造现实世界的成本和门槛，而其中人类语言（又称自然语言）具有歧义性、抽象性、无穷的语义组合性和持续进化性等特点，并且理解语言往往需要具有一定的知识推理和认知能力，因此自然语言处理领域是人工智能技术突破的关键难点，是制约人工智能取得更大跃升和更广泛应用的瓶颈之一，又被誉为“人工智能皇冠上的明珠”[1]。自 2022 年底生成式人工智能 AIGC 技术的爆发式增长已突破了这一障碍，并让全球新闻舆论格局首当其冲面临了颠覆式的改变。

二、ChatGPT 开启生成式人工智能（AIGC）时代

美国 OpenAI 公司从 2018 年起开始专注于 GPT 系列大规模生成式预训练语言模型的技术路线，在“大规模数据+大规模算力+大规模参数=大模型”基础上探索出了“基础大模型+指令微调”的人工智能新范式[2]，突破了人工智能理解处理和生成自然语言的瓶颈。基于大规模预训练语言模型 GPT-4 的应用 ChatGPT，可以通过与人类进行多轮对话的方式，识别人类意图和隐喻、理解对话上下文、

进行逻辑思考和推理、生成内容完整清晰合理的回答、优化内容中的知识点和措辞风格，并可以进一步通过接口对接集成到各类应用程序中，扩展执行多类任务，涌现出了不同以往的智能水平，展现了如下能力：

1、具备通用知识水平并可向不同专业领域扩充和掌握知识。通过增加专业领域的训练数据和多个领域专家大模型之间的配合，可以扩展解决多种复杂问题；

2、具备联想和创作能力。创造隐喻并挖掘事物之间关联，甚至可以理解幽默和生成段子、诗歌与小说；

3、具备思维链推理能力。可以自行将需要逻辑推理的复杂问题拆解成步骤，逐步给出解答过程和答案；

4、具备抽取和总结知识与主要观点的能力。可以将长文章中的内容摘要、大纲、知识点抽取生成出来；

5、具备根据需求自动生成和检查程序代码的能力。可以根据设计图和需求描述生成可以执行的程序代码。

微软日前发表论文称对 GPT-4 进行了全面评测，认为“鉴于 GPT-4 能力的广度和深度，它应该被合理视作一个通用人工智能（AGI）系统的早期（但仍不完整）版本”[3]。GPT-4 及其应用 ChatGPT 标志着人工智能从感知理解世界进入到了生成创造世界的新阶段。

三、高质量训练数据集是 AIGC 的关键

从 GPT-1 到 GPT-4 的大模型进化过程中，除了算力基础设施外，高质量大规模数据集是决定大模型能力的关键因素，根据 OpenAI 前期论文和博客介绍，ChatGPT 中数据集的规模和构建质量均高于以往的人工标注数据集[4]，ChatGPT 大模型采用的 Transformer 架构解码预训练模型的原理本质上是通过数据集语料中字词出现的概率和关联关系来抽取特征，在已有字词后面预测补充最有可能出现的字词来实现语言理解和生成的，因此训练数据集的收集、清洗和特定标注异常重要：

首先、GPT-4 的基础预训练是在大量无标注、但需要质量高、重复率少、噪音小、知识密度高、规范化程度高的大规模数据集上进行自监督训练来完成的，保证大模型具备正确的语言理解和生成能力，训练数据集包括 13 万亿 token（单词或字符）的语料，涵盖全球互联网中主要以西方发达国家平台为主的数据源，如维基百科、电子书籍、科学期刊、reddit 社交媒体点赞数多的评论数据集、commonCrawl 网页数据集等。

其次、ChatGPT 的大规模预训练语言模型 GPT-4 还通过大量来自 GitHub 的开源程序代码数据集、代码注释数据约 4.5TB，这部分面向具体问题和需求、有结构化分解和实现步骤注释的代码数据让 GPT-4 拥有了思维链（COT）能力和部分逻辑推理能力。

第三、GPT-4 基础预训练模型还需经过人工调优以及用带有人工标注的数据集进行有监督训练，一方面适应不同专业领域的问题，正确理解任务需求，生成更准确合理的内容；一方面实现与人类意图对齐，即判别人类恶意指令、按照人类指令尽可能生成无负面影响结果的内容。这类数据集分为两大类，一类是提示学习和指令精调数据集，主要有一系列问答对，提示指令、问题集及对应的相关内容文本语料构成；一类是用于进行 RHLF（人类反馈强化学习）的数据集，请专家对大模型按照指令给出的答案和内容进行打分，标注人类偏好标签，通过奖励

模型训练，让算法拟合人类的期望和倾向，减少有害内容，优化大模型的参数策略。[5]

从上述预训练语言模型的训练原理可以看出，大规模数据集让 AIGC 大模型掌握了人类公开在互联网上的大量知识和原创内容，赋予了人工智能类人类的对话交互能力、知识体系和思考分析过程，而 ChatGPT 通过这样的自然语言入口，依托大模型快速构建起了应用生态，一是以 ChatGPT 接口能力，在教育、传媒、商务、客服、办公、内容出版等领域成为人类进行内容创作和生成的得力助手，二是类 GPT-4 的 AIGC 大模型通过补充专业领域数据集和语料集，让构建医疗、制造、交通、法务、政务、汽车制造等产业端行业 AI 基础服务的成本和难度大大降低，加速产业数智化转型和高质量发展；三是 AIGC 大模型开始提供应用程序插件功能，形成了用人类自然语言操作各类应用程序完成任务的总入口，基于 AIGC 大模型能力的进一步提升，结合应用程序插件，可以自行寻找链接程序接口和数据源的 AI Agents（智能体）研究将成为 OpenAI 的下一个研究突破的目标，AI Agents 可以根据人类一句任务指令，自行分析、分解、优化，进化出解决任务的能力，并寻找合适资源完成任务。[6]

四、AIGC 时代新闻舆论格局面临的风险与挑战

AIGC 大模型的特性和应用生态的发展趋势预示着以大模型和内容为核心驱动的新一代数字经济形态正在逐步形成，模型即服务成为数智化转型的服务载体，自然语言成为人机交互的指令载体，而内容数据本身作为大模型训练必备的数据集及语料，又是 AIGC 大模型生成的重要形态，其作用从以往的信息载体向知识载体甚至是生产力载体进化，内容生产传播体系与社会经济生活的运行正前所未有的深度融合绑定。

人工智能发展的每一个阶段都会推进和影响社会意识形态或主流价值观的塑造方式，为新闻舆论工作提供新的平台和模式。 物联网、大数据、云计算、区块链、算法系统在网络空间中构建出独特的公共舆论体系，以网络平台为新闻舆论聚集地和扩散源，将公众汇集成各种不同的价值群体和多元的意识形态群体[7]，其中推荐算法控制了内容传播的范围和可见度；而 AIGC 大模型的出现让数据集和原创内容成为人工智能感知现实世界获取知识的媒介、成为内容生产的关键要素，算法和算力逐步掌握内容生产和传播的权力核心，随着内容驱动的数字经济生态不断丰富，AIGC 大模型成为潜在的社会舆论成员，并以远超人类个体的知识面和内容处理生成速度潜移默化的掌握了舆论引导的主动权和话语权，在主流媒体新闻舆论场、新兴自媒体新闻舆论场上又叠加了生成式人工智能大模型新闻舆论场，迫使当前新闻舆论工作从“生产端”、“流通端”到“作用端”的构建方式与运行机制发生改变。

改变的核心一方面是要把 AIGC 大模型这样的人工智能纳入到工作全流程来考虑；另一方面要重视内容驱动下舆论场与社会政治经济文化生活方方面面的深度融合。新闻舆论工作不仅要做好主流媒体与新兴自媒体间的协调联动，还要做好与人工智能 AIGC 大模型之间的协调联动；不仅要做好面向人的新闻舆论工作，还要做好面向人工智能的新闻舆论工作。由于影响 AIGC 大模型能力的关键因素是内容数据集，且对实际社会经济生活产生作用的中介也是内容数据，因此面向 AIGC 大模型训练的内容数据集和数据服务建设是新闻舆论工作必须高度重视的

阵地。尤其当下美西方国家人工智能巨头如OpenAI、Meta、Google等陆续推出的AIGC大模型不断成为各行各业人工智能应用发展的基座，会给我国主流新闻舆论格局带来诸多风险与挑战：

首先，AIGC高仿真内容生成导致虚假新闻泛滥：AIGC大模型有着高度逼真的内容生成能力，其语言逻辑通顺、图像逼真清晰，会出现捏造答案和伪造事实的现象，且生产和传播速度极快，导致虚假信息泛滥。如美国媒体机构G/O Media在旗下的科技网站Gizmodo上，使用谷歌Bard和OpenAI的ChatGPT编写了一篇有关《星球大战》的文章，出现了诸多事实错误；科大讯飞也因为AI自动生成关于“涉嫌大量采集用户隐私数据”、“美国正在考虑是否将科大讯飞、美亚柏科等加入制裁名单”的假消息导致股价闪崩。

其次，AIGC的内容生成机制难以解释和追溯让舆论溯源更困难：AIGC大模型是通过概率模型参数逐字推测来实现内容生成，算法黑盒导致难以解释和溯源，生成内容具有随机性和无法复现的问题，缺少时效性和时序性，观点、事实、知识的来源无法查证，使得真相与虚假杂糅同构[8]，对于AIGC生成的议题设置、舆论观点、伪事实内容和内容侵权，若无人工审核校验留痕，都很难进行源头追溯和传播追踪。

第三，人机对话点对点交互方式让舆论发现和引导更被动:AIGC大模型通过与人类对话的方式进行交互和内容输出，舆论引导和传播从公域转向了点对点的私域；人工智能在深度学习中对大量用户敏感数据的交互使用，不仅使人类隐私暴露在人工智能之下，也极大地削弱了政府对数据信息的监管能力[9]。信息传播的高度个性化和即时性可以更深入的影响用户的认知，在公域互联网空间内越来越难掌握到真正公众对事件的舆情动向、意见看法、信念态度，难以有针对性的进行解读、引导并促进舆情化解和达成共识。而AIGC大模型是否有正确的引导力完全有赖于大模型训练和优化所使用的数据集和人工智能训练专家。

第四，AIGC的技术霸权属性让舆论操控更隐蔽：虽然OpenAI创始人认为AIGC人工智能可以帮助人们快速掌握知识，提升能力，让知识资源更平等服务于每个人。但实质上AIGC大模型依赖的是庞大的算力和数据集，在使用过程中又不断的将人类原创内容和智慧甚至隐私信息吸收到掌握大模型技术和服务的机构中，占据技术创新优势的美西方国家以及有足够资本支撑大规模算力和数据集生产高昂成本的机构通过技术霸权成为了舆论话语权的隐形垄断者，通过收集个人信息，通过大数据进行群体画像分析，或许会成为大模型掌控者研究和制定思想渗透策略的重要数据支撑[10]，通过AIGC全方位影响和塑造用户的知识领域、意识形态和价值判断，进而形成认知茧房，形成舆论操控的超级中心化。

第五，AIGC带有较难扭转的价值观和意识形态属性让影响舆论更为深远：AIGC大模型的训练方式决定了人工智能不仅学会了自然语言的文法和表述方式，还抽取和学习到了知识、立场、观点和价值判断，AIGC大模型带来的不仅是信息的传播，更需要警惕的是带来了意识形态和价值观的传播，AIGC大模型内在价值观一旦形成很难完全扭转和改变，如ChatGPT的价值观底色根植于参与该系统设计研发人员的价值观取向[11]，取决于集中体现美西方意识形态和价值观判断的书籍、百科、社群讨论和网站。而ChatGPT的迅速流行会使用户产生依赖进而削弱批判思维的形成和接触现实的机会，因此缺少自主训练数据集的大模型广泛应用必将对我国主流新闻舆论格局造成更大冲击。

世界各国也都意识到了AIGC对国家秩序、社会伦理、舆论空间的风险与影响。美国国家标准与技术研究院发布人工智能风险管理框架，美国计算机协会的

全球技术政策委员会也发布了《生成式人工智能技术的开发、部署和使用原则》；意大利个人数据保护局率先封禁了 ChatGPT，法国、爱尔兰、德国等国也跃跃欲试考虑采取封禁措施，担忧技术失控的情绪正在全球蔓延[12]。2023 年 8 月 15 日国家网信办联合六部委发布的《生成式人工智能服务管理暂行办法》正式施行，而这些监管规则有效落地实施，需要一个共性基础条件，就是面向人工智能的可信训练数据集和数据服务能力建设。

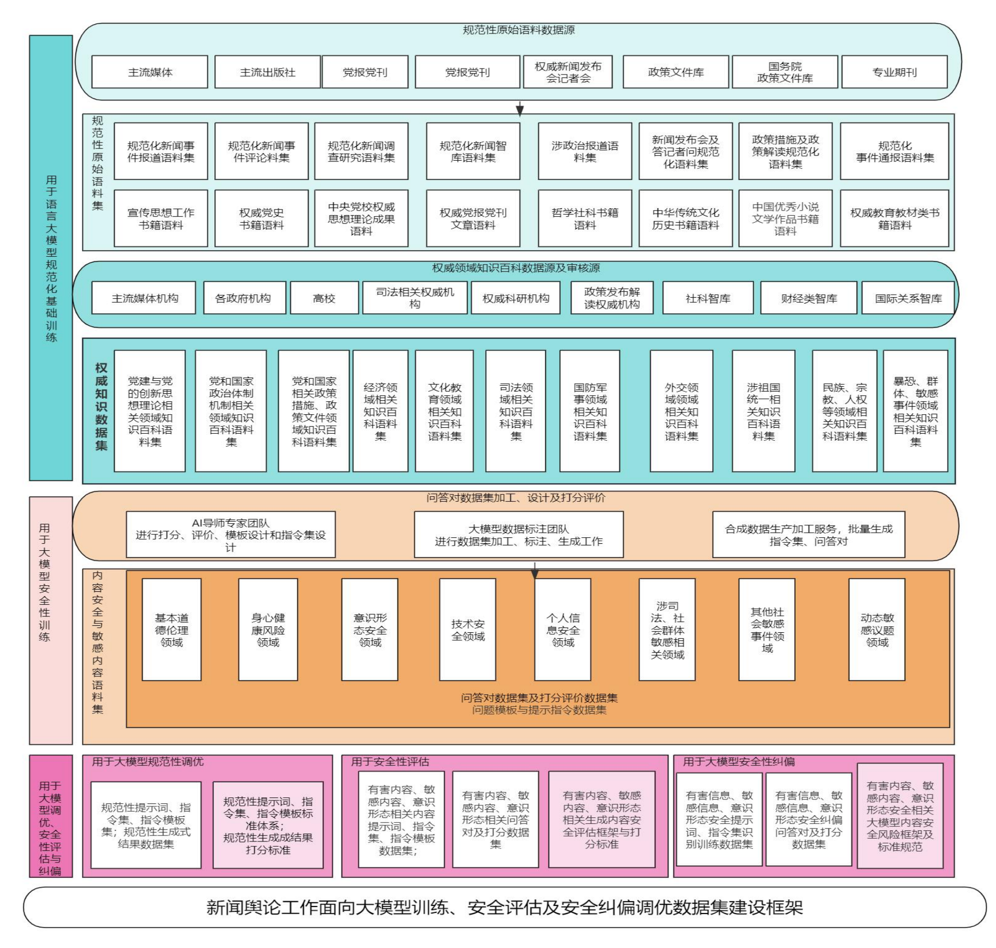
五、进军新闻舆论工作新阵地——可信训练数据集及数据服务

当前我国 AIGC 大模型研发风生水起，截止 7 月份，已发布通用大模型和行业大模型 100 余个，10 亿参数规模以上的为 79 个，囿于奇高的算力成本和带有中国主流价值观和意识形态的高质量训练数据语料集的缺乏，大多数中国的大模型还是在美西方开源大模型基础上进一步训练调整而来，同时西方国家的科研团队也在抓紧进一步挖掘中文领域训练数据集的富矿，如近期 Meta 的 AIGC 大模型 Llama 2 的合作伙伴中包括了中国 AI 训练数据提供商海天瑞声，并共同发布了超大规模中文对话数据集 DOTS-NLP-216。

党的新闻舆论工作涉及到“五个事关”，责任意义重大，中国主流新闻舆论工作者肩负着为大众提供真实新闻信息、引导和监督舆论的职责，承担着发挥“舆论压舱石、社会黏合剂、价值风向标”、“构建网上网下一体、内宣外宣联动的主流舆论格局”的使命。在人工智能发展带来的风险挑战和严峻形势下，主流媒体新闻舆论工作者如何“探索将人工智能运用在新闻采集、生产、分发、接收、反馈中，用主流价值导向驾驭‘算法’，全面提高舆论引导能力”，最重要的是充分发挥主流新闻舆论工作者脚力、眼力、脑力、笔力积累，恪守新闻伦理和社会责任的专业素养，把握处于 AI 上游通过调查研究接触现实世界一手资料的优势地位，面向人工智能 AIGC 大模型不仅做到“守土有责”，做好“把关人”角色，更要做到“开疆扩土”，开辟面向大模型训练的可信数据集和数据服务新阵地，提供决定大模型核心能力和价值观的内容供给与知识供给，抢占 AIGC 时代舆论引导、思想引领、文化传承、服务人民的传播高地。

新闻舆论领域提供的可信训练数据集与数据服务建设包含三层含义：一是内容数据规范权威真实；二是内容数据可溯源可确权；三是符合主流价值且可审核可纠偏。围绕这三层含义需开展如下建设内容：

首先，建立 AIGC 大模型全生命周期训练数据集：包含四大类，一是建立高质量规范化数据集和语料集，充分覆盖主流意识形态和价值观的规范化表述，包括高质量书籍，权威解读，标准问答，新闻事实稿件、述评和调查研究，保证大模型语言、立场、观点和思维方式的准确性、规范化与专业性；二是建立保证事实与知识准确性的高质量领域知识库数据集，尤其涉及中国政治、社会、经济、文化等领域的权威阐述。三是建立内容意识形态安全语料集和主流价值观语料集，主要有涉及意识形态安全的问题与指令集，问答对，问答模板以及评价打分数据集，用于对基础大模型进行价值观与意识形态纠偏和对齐；四是建立用于保证 AIGC 在多场景下生成内容的规范性评估、安全性评估和纠偏数据集，包括大模型规范性评估、有害内容与敏感内容检查评估、意识形态纠偏所需的指令集、指令模板、提示词、打分数据集和问答对数据集。



新闻舆论工作面向大模型训练、安全评估及安全纠偏调优数据集建设框架

其次，建立相关审核打分和大模型意识形态与价值观评价标准规范：大模型训练数据集建设还需要配套相关标准规范，包括基础训练数据清洗去重标注规范；知识库知识框架和审核规范；指令集、指令模板、问答对、提示词标注标准规范；指令模板和提示词规范以及一系列人类专家反馈强化学习打分与标签标准规范；技术伦理、有害内容、敏感内容的分类分级标准规范等。

第三，建立主流大模型人工标注与专家反馈合作服务机制：形成面向大模型的常态化专家训练合作机制和面向社会提供专家训练服务的机制，输出代表中国权威知识内容和主流意识形态的专家智慧。一是组织国际关系、社会科学、新闻传播等领域的学生和从业者构成主流大模型训练数据集标注和指令集生成团队；二是组织各领域学界权威专家、智库学者和知识内容原创者形成知识库内容审核团队，确保知识体系框架正确，内容表述准确完整；三是组织新闻舆论和传播领域资深专家、智库学者形成大模型人类反馈强化学习的 AI 导师团队，构建人类反馈强化学习数据集，开展大模型意识形态审核和评估；四是逐步依托主流大模型提供合成数据生成服务，通过主流大模型本身大规模生成主流意识形态训练数据集，有效弥补领域数据量不足的问题，提升数据集生产和标注效率。

第四，建立动态追踪和审核大模型意识形态安全服务：形成面向国内外大模型的意识形态安全动态追踪和审核机制，为即将推出服务和已经开展服务的 AIGC 大模型提供上线前内容安全审核评估服务、上线后内容安全追踪服务，动态收集各类内容安全事件、安全问题、不断丰富补充主流大模型所需的评估审核数据集，同时有针对性的丰富完善大模型意识形态安全纠偏训练数据，为大模型的各类商业应用提供内容安全修正和优化服务。

第五，建立适应 AIGC 大模型的数据安全、内容追溯和事实核查机制：AIGC 大模型训练数据集涉及到数据源、内容原创者、使用者等多方利益，也存在数据安全、隐私保护和数据真实性问题，需要面向安全可信、隐私保护、版权追溯的需求创新训练数据集生产和服务的技术手段、平台工具、加工流程和标准规范，支持多方安全计算和联邦计算方式，支持安全可控可追溯可确权的人工智能模型训练需求；形成主流新闻舆论工作者在 AIGC 大模型研发、服务、融合应用各环节做好内容安全和事实核查把关人的机制。

六、结语：

新一代人工智能发展趋势下，我国新闻舆论工作必须将人工智能作为新的舆论主体纳入到新闻舆论工作流程再造中来，深刻认识人工智能时代新闻舆论工作中“四力”核心竞争力的重要意义，并将其转化为面向大模型的训练数据集和内容供给，快速占领 AIGC 上游新高地，深度融合到社会经济运行场景中，一方面充分运用 AIGC 技术延伸主流新闻舆论工作效能，推动多元话语体系互动融合，构建新型舆情态势感知、应对、引导模式；一方面为 AIGC 技术伦理约束与技术监管落地提供强有力的内容、机制和服务保证。

参考文献：

[1] 车万翔，刘挺. 自然语言处理新范式：基于预训练模型的方法[J]. 中兴通讯技术. 2022,28(02):3-9.

[2] 张飞飞,张禹,徐才,柴青山. 6G——搭建多维感知与信息融合网络 中国传媒科技. 2023(03):19-22.

[3] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, Yi Zhang. Sparks of Artificial General Intelligence: Early experiments with GPT-4. ArXiv preprint, 2023, abs/2303.12712.

[4] OUYANG L, WU J, JIANG X, et al. Training language models to follow instructions with human feedback[J].ArXiv preprint, 2022, abs/2203.02155.

[5] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, Ji-Rong Wen.A Survey of Large Language Models.arXiv:2303.18223.

[6] Joon Sung Park, et al. “Generative Agents: Interactive Simulacra of Human Behavior." arXiv preprint arXiv:2304.03442 (2023).

[7] 张爱军,贾璐.ChatGPT 的公共政治舆论情感驱动及其调适[J].探索，2023(3):162-174

[8] 王延川,赵靖.生成式人工智能诱发意识形态风险的逻辑机理及其应对策略[J], 河南师范大学学报(哲学社会科学版).2023,50(04):1-7

[9] 黄日涵,姚浩龙.被重塑的世界？ChatGPT 崛起下人工智能与国家安全新特征. 国际安全研究. 2023,41(04):82-106+158-159

[10] 李御任,翟红蕾. ChatGPT 模型的舆论风险及应对策略研究.视听. 2023(05):7-10

[11] 陈昌孝,王梓晗.认知视角下 ChatGPT 的特征、运用及应对之策.政工学刊. 2023(07):59-61
[12] 张欣.生成式人工智能的数据风险与治理路径. 法律科学(西北政法大学学报). 2023(05):1-13